Assignment2 - Step 3

Phoebe Zhou

September 2025

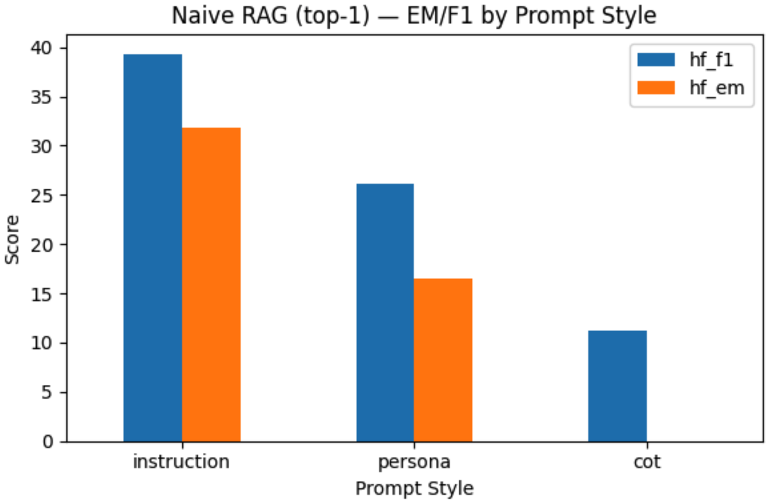
Evaluation Phase I — Initial Results Report

For this phase, I evaluated the RAG pipeline using three prompting strategies: Instruction Prompting, Persona Prompting, and Chain-of-Thought Prompting. Each question was answered using only the top-1 retrieved passage from the RAG index, and answers were evaluated against the gold standard using Exact Match (EM) and F1 score, implemented with Hugging Face's Squad metric.

1. Results Overview

The results reveal clear differences between prompting styles:

- Instruction Prompting achieved the best performance, with an F1 score of 39.3 and EM score of 31.8 across 918 questions.
- Persona Prompting yielded an F1 of 26.1 and EM of 16.6.
- Chain-of-Thought Prompting performed worst, with F1 11.2 and EM essentially 0, indicating it struggled to produce answers aligned with gold references.



The bar chart visualization makes this contrast particularly stark: instruction prompts nearly double the accuracy of persona prompts, and vastly outperform CoT prompts in this naïve setup.

2. Interpretation

The dominance of instruction prompting is not surprising. The instruction template explicitly guides the model to use the retrieved passage to generate concise answers. This directness likely minimizes hallucinations and keeps the answer closer to the gold standard wording, boosting EM and F1.

Persona prompting added extra style and explanation to the answers. While this made the responses sound friendlier and more conversational, it often strayed from the exact wording of the gold answers. As a result, the EM score dropped because the model's answers didn't match the reference text closely enough.

The poor performance of Chain-of-Thought prompting in this setup is revealing. While CoT is often powerful for reasoning-heavy tasks, here the dataset consists of factual QA pairs with direct answers available in short passages. CoT encouraged the model to generate verbose reasoning steps, which strayed from the concise ground truth. As a result, answers mismatched the expected labels, especially under strict EM scoring.

3. Hypothesis on Best Strategy

From this evaluation, Instruction Prompting emerges as the best strategy for factual QA in a RAG context where answers are expected to match reference spans. Its balance of guidance and conciseness allows the model to leverage retrieved evidence effectively while staying close to gold-standard wording.

Going forward, I might refine persona or CoT prompts to better constrain output. However, in their current form, both reduce alignment with ground truth. A hybrid approach (e.g., CoT reasoning internally but truncating output to a concise final answer) could close this gap.